15-851 Algorithms for Big Data — Spring 2025

Problem Set 1

Due: Thursday, February 6, before class

Please see the following link for collaboration and other homework policies:
http://www.cs.cmu.edu/afs/cs/user/dwoodruf/www/teaching/15851-spring25/grading.pdf

Problem 1: Subspace Embeddings via Random Sign Matrices (17 points)

In class we showed that if $k = O(d/\epsilon^2)$ and we choose a random $k \times n$ Gaussian matrix S so that each entry is i.i.d. $N(0, 1/k)$, then with probability at least $9/10$, we have simultaneously for all x that $\|SAx\|_2^2 \in (1 \pm \epsilon)\|Ax\|_2^2$.

Now suppose we instead choose a $k \times n$ matrix S where each entry is independently chosen to be $+\frac{1}{\sqrt{k}}$ with probability $1/2$, and chosen to be $-\frac{1}{\sqrt{k}}$ with probability $1/2$. In this problem we will show for appropriate $k = O(d/\epsilon^2)$ that we again have with probability at least $9/10$, simultaneously for all x that $\|SAx\|_2^2 \in (1 \pm \epsilon)\|Ax\|_2^2$. We prove this in steps:

1. (2 points) Show that for any fixed $x \in \mathbb{R}^d$, we have $\mathbf{E}_S[\|SAx\|_2^2] = \|Ax\|_2^2$.

   The above part shows that we are correct in expectation for a fixed x. We next need to understand the deviation of $\|SAx\|_2^2$ from its expectation, for which we study the tail behavior of random variables.

2. (3 points) A zero-mean random variable Y is sub-Gaussian with parameter σ^2 if $\mathbf{E}[e^{tY}] \leq e^{\sigma^2 t^2/2}$ for all t. Argue that if $Y \in \{-1, 1\}$ is chosen uniformly at random, then Y is sub-Gaussian with parameter $\sigma^2 = 1$.

   HINT: One can use properties of $\cosh(t)$ to prove this, or one can use the Taylor series $e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots$ a few times and compare terms.

3. (2 points) If $Y_1, \ldots Y_n$ are independent zero-mean σ^2-sub-Gaussian random variables, then for scalars $\alpha_1, \ldots, \alpha_n$, show that $Y = \sum_i \alpha_i Y_i$ is $\sigma^2 \cdot \sum_i \alpha_i^2$-sub-Gaussian.

4. (3 points) In this part we will use the following fact, which you can use without proof and follows by direct integration: for $V \sim N(0, \sigma^2)$, $\mathbf{E}[e^{tV}] = e^{t^2\sigma^2/2}$.

   Now suppose Y is mean-zero σ^2-sub-Gaussian, and also suppose Y is symmetric around the origin. Prove that for $V \sim N(0, \sigma^2)$, for any $t > 0$ that

   $$\mathbf{E}[e^{tY^2}] \leq \mathbf{E}[e^{tV^2}].$$

   HINT: Start by arguing that $\mathbf{E}_Y[e^{tY^2}] = \mathbf{E}_{Y,V}[e^{(\sqrt{2t}|Y|/\sigma)V}]$ using the fact above, then use the fact that both Y and V are symmetric.

5. (5 points) Using parts 2-4 above, argue that for appropriate $k = O(d/\epsilon^2)$ that for any fixed $x \in \mathbb{R}^d$,
$$\Pr[|\|SAx\|_2^2 - \|Ax\|_2^2| \geq \epsilon \|Ax\|_2^2] \leq e^{-\Theta(d)}.$$
HINT: $Y = \|SAx\|_2^2$ is an average of squares of k independent symmetric sub-Gaussian random variables Y_i, for $i = 1, \ldots, k$. For the upper tail bound, start by writing
$$\Pr[Y \geq 1+\epsilon] = \Pr[e^{tkY} \geq e^{tk(1+\epsilon)}] \leq \frac{\mathbf{E}[e^{tkY}]}{e^{tk(1+\epsilon)}} = \prod_{i=1}^{k} \frac{\mathbf{E}[e^{tY_i^2}]}{e^{t(1+\epsilon)}},$$
which holds for any $t > 0$, and where the inequality is by Markov's bound. Then use your result from part 4. You can also use the fact that for $V \sim N(0,1)$ and $t < 1/2$, that $\mathbf{E}[e^{tV^2}] \leq \frac{1}{\sqrt{1-2t}}$, which follows by direct integration. You might also need to expand a Taylor series to derive a tractable tail bound.

For the lower bound, you can start in a similar way. Then you can use the following derivation. If Y is mean zero subgaussian with parameter $\sigma^2 = 1$, then the following is true for $|t| < 1$. Using the Taylor expansion, we have $E[e^{tY^2}] \leq 1 + tE[Y^2] + t^2 \sum_{i \geq 2} E[Y^{2i}/i!]$. Now since we know $E[Y^2] = 1$ and $|t| < 1$, this is at most $1 + t + t^2 E[e^{Y^2}]$. Now notice that $E[e^{Y^2}]$ is part of the upper tail, and so we get that $E[e^{tY^2}] \leq 1 + t + t^2/\sqrt{1-2t}$.

6. (2 points) Conclude that for appropriate $k = O(d/\epsilon^2)$ that with probability at least $9/10$, simultaneously for all x, we have $\|SAx\|_2^2 \in (1 \pm \epsilon)\|Ax\|_2^2$. You are welcome to cite anything from class without proof.

Problem 2: Multiplying Gaussian Matrices (10 points total)

Let g_1 and g_2 be standard $N(0,1)$ Gaussian random variables. Note that $g_1 \cdot g_2$ is not a Gaussian random variable. We can ask a similar question for matrices. Suppose we have a $d \times t$ matrix G_1 of i.i.d. $N(0,1)$ entries and a $t \times d$ matrix G_2 of i.i.d. $N(0,1)$ entries where $t = \omega(d^2)$ ($\lim_{d \to \infty} \frac{t}{d^2} = \infty$) and we look at the $d \times d$ matrix $G_1 \cdot G_2$. In this problem you will prove that $G_1 \cdot G_2$ cannot be distinguished from a $d \times d$ matrix H of i.i.d. $N(0,t)$ random variables.

To make the above statement precise, we will use a result of Jiang which states the following: let A be an arbitrary, possibly randomized algorithm. Consider an $r \times \ell$ submatrix X of a random $z \times z$ matrix with orthonormal rows and columns. We refer to the distribution of X as p. Also, consider an $r \times \ell$ matrix Y with i.i.d. $N(0, 1/z)$ entries. We refer to the distribution of Y as q. Suppose with probability $1/2$ we give a random sample from p to algorithm A, while with the remaining probability $1/2$ we give a random sample from q to algorithm A. If we have $r \cdot \ell = o(z)$, then the probability that A correctly states if its input was chosen from p or from q is at most $1/2 + o(1)$, where $o(1) \to 0$ as $z \to \infty$. This says that small submatrices of random orthonormal matrices are indistinguishable from

Gaussian matrices. Intuitively, one cannot "observe" the orthonormality constraints on a small submatrix of a random orthonormal matrix.

Using the above result, we will prove the following. If A is an arbitrary, possibly randomized algorithm where p' is the distribution of $G_1 \cdot G_2$ and q' is the distribution of H, then if we randomly give A a sample from p' with probability $1/2$ while with the remaining probability $1/2$ we give a random sample from q', then the probability that A correctly states if its input was drawn from p' or q' is at most $1/2 + o(1)$.

1. (2 points) Write $G_1 = U\Sigma V^T$ (in its SVD) and consider $U\Sigma V^T G_2$. Show that $V^T G_2$ is a $d \times d$ matrix of i.i.d. $N(0,1)$ entries.

2. (2 points) Now take $M = V^T G_2$. Using Part 1, show that M is indistinguishable from $\sqrt{t} \cdot \tilde{M}$ where $\tilde{M}$ is a $d \times d$ submatrix of a random matrix with orthonormal rows and columns.

   HINT: Use Jiang's result.

3. (6 points) Using Part 2, show that the probability A correctly states if its input was drawn from p' or q' is at most $1/2 + o(1)$.

   HINT: Think about writing $\tilde{M}$ as a product of two other matrices. It will be helpful and you can freely use the fact that the SVD of a random $d \times t$ matrix G_3 of i.i.d. $N(0,1)$ random variables is equal to $U\Sigma V^T$, where $U, \Sigma \in \mathbb{R}^{d\times d}$ and $V^T \in \mathbb{R}^{d\times t}$ are independent matrices and V^T is a random matrix with orthonormal rows.

Problem 3: Learning the Positions and Values of CountSketch (10 points)

In class we claimed that if S is an $m = O(d^2/(\epsilon^2\delta)) \times n$ CountSketch matrix, then for any fixed $n \times d$ matrix A, we have that with probability at least $1 - \delta$, simultaneously for all x,

$$\|SAx\|_2^2 = (1 \pm \epsilon)\|Ax\|_2^2.$$

The number m of rows in CountSketch may be too large for some applications. Recall that CountSketch is constructed randomly, i.e., for each column we independently choose a non-zero location uniformly at random and place $+1$ in that location with probability $1/2$, and -1 in that location with probability $1/2$.

To try to improve the number of rows in S, one can try to *learn* the best location in each column to place a non-zero entry, as well as the best value to put in the non-zero location in each column of S. Note that S will still only have a single non-zero entry per column, but the location of this entry need no longer be random and its non-zero value can be arbitrary.

Suppose one is given as input an $n \times d$ input matrix A for which each row of A has only a single non-zero entry. Design a deterministic matrix S of the form described in the previous paragraph, which may depend on A, so that S has exactly d rows and $\|SAx\|_2^2 = \|Ax\|_2^2$ for all x.

Problem 4: Approximate Matrix Product in Terms of Stable Rank (13 points)

In class we saw an approximate matrix product lemma, namely, given an $n \times d$ matrix A and an $n \times e$ matrix B, for certain random families of matrices S with $O((\log n)/\epsilon^2)$ rows:

$$\Pr_S[\|A^T S^T SB - A^T B\|_F^2 \geq \epsilon^2 \|A\|_F^2 \|B\|_F^2] \leq \frac{1}{\mathrm{poly}(n)}.$$

The error in terms of the Frobenius norm can be large, so an alternative desirable guarantee could be to design a random family of matrices S with a small number of rows for which:

$$\Pr_S[\|A^T S^T SB - A^T B\|_2^2 \geq \epsilon^2 \|A\|_2^2 \|B\|_2^2] \leq \frac{1}{\mathrm{poly}(n)}, \tag{1}$$

where for a matrix C, we have $\|C\|_2 = \sup_{x \neq 0} \frac{\|Cx\|_2}{\|x\|_2}$ is its operator norm. For ease of notation, let us assume $d = e$ in the remainder of this problem.

1. (4 points) Give an example for which $A = B$ and for $\epsilon = 1/2$ for which any such family S of matrices which satisfies Equation 1 would require $\Omega(d)$ rows.

   HINT: Consider the case when $n = d$ and $A = B = I$. Then generalize this to $n \neq d$.

2. (2 points) While the previous part shows that for worst case matrices A and B the number of rows of S needs to grow linearly with d in order to achieve (1), in many practical cases we can do better. The *stable rank* $\mathrm{srank}(A)$ of an $n \times d$ matrix A is defined as $\frac{\|A\|_F^2}{\|A\|_2^2}$. Argue that $\mathrm{srank}(A) \leq d$ for any $n \times d$ matrix A.

   HINT: Take the singular values of A to be $\sigma_1 \geq \sigma_2 \geq \sigma_3 \ldots$. You can use the fact that $\|A\|_2 = \sigma_1$ and $\|A\|_F = \sqrt{\sum \sigma_i^2}$.

3. (7 points total) We now prove an approximate matrix product lemma, which shows that if S has $m = O((\epsilon^{-2} \log n)(\mathrm{srank}(A) + \mathrm{srank}(B)))$ rows and corresponds to a random sampling and rescaling matrix from a distribution described below, then we can achieve (1). Note that the number of rows of S can be significantly smaller than d, as the stable ranks of A and B could be constant in typical applications. We will use a generalization of the Matrix Chernoff lemma from class, which you can use without proof:

   (**Generalized Matrix Chernoff**) Let F be a fixed $d \times d$ matrix and suppose R is a random matrix with $\mathbf{E}[R] = F$ and $\|R\|_2 \leq L$ with probability 1, for a parameter L. Let $\beta_2(R) = \max(\|\mathbf{E}[R^T R]\|_2, \|\mathbf{E}[RR^T]\|_2)$ and let $\bar{R}_m = \frac{1}{m}\sum_{i=1}^m R_i$ where each R_i is an independent copy of R. Then for every $t > 0$ we have:

   $$\Pr[\|\bar{R}_m - F\|_2 > t] \leq 2d \cdot \exp\left(\frac{-mt^2/2}{\beta_2(R) + 2Lt/3}\right).$$

Returning to our problem, let $p \in [0,1]^n$ be any probability distribution such that for all $i \in \{1, \ldots, n\}$:

$$p_i \geq \frac{1}{4} \cdot \frac{\|A_i\|_2^2 + \gamma\|B_i\|_2^2}{\|A\|_F^2 + \gamma\|B\|_F^2},$$

where $\gamma = \|A\|_2^2/\|B\|_2^2$ and A_i and B_i are the i-th row of A and B, respectively. Suppose we create the sampling and rescaling matrix $S \in \mathbb{R}^{m\times n}$ by first generating m samples $\ell_1, \ldots, \ell_m$ with replacement from p, and then letting the i-th row of S equal $\frac{1}{\sqrt{mp_{\ell_i}}} \cdot e_{\ell_i}^T$, where $e_{\ell_i}^T$ is the ℓ_i-th standard (row) unit vector. We will show that for $m = O((\epsilon^{-2}\log n)(\mathrm{srank}(A) + \mathrm{srank}(B)))$, (1) holds.

(a) (1 point) Determine what R_i for $i \in \{1, \ldots, m\}$ is and show that $A^T S^T S B = \bar{R}_m = \frac{1}{m}\sum_{i=1}^m R_i$.

(b) (1 point) Show that $\mathbf{E}[R] = A^T B$.

(c) (2 points) Show that $L = O(\|A\|_2\|B\|_2(\mathrm{srank}(A) + \mathrm{srank}(B)))$.

HINT: You can use the AM-GM inequality which says for two nonnegative numbers x and y, we have $\frac{x+y}{2} \geq \sqrt{xy}$.

(d) (2 points) Show that $\beta_2(R) = O(\|A\|_2^2\|B\|_2^2(\mathrm{srank}(A) + \mathrm{srank}(B)))$.

(e) (1 point) Conclude that S with $m = O((\epsilon^{-2}\log n)(\mathrm{srank}(A)+\mathrm{srank}(B)))$ satisfies (1).